UNIVERSITY OF ILLINOIS BULLETIN

ISSUED WEEKLY

VOL. XXVI MAY 7, 1929 No. 36

[Entered as second-class matter December 11, 1912, at the post office at Urbana, Illinois, under the Act of August 24, 1912. Acceptance for mailing at the special rate of postage provided for in section 1103, Act of October 3, 1917, authorized July 31, 1918.]

BULLETIN NO. 46

BUREAU OF EDUCATIONAL RESEARCH
COLLEGE OF EDUCATION

THE USE OF SCALES FOR RATING PUPILS' ANSWERS TO THOUGHT QUESTIONS

By

C. W. ODELL
Assistant Director, Bureau of
Educational Research

PRICE 50 CENTS

PUBLISHED BY THE UNIVERSITY OF ILLINOIS, URBANA
1929

The Bureau of Educational Research was established by act of the Board of Trustees June 1, 1918. It is the purpose of the Bureau to conduct original investigations in the field of education, to summarize and bring to the attention of school people the results of research elsewhere, and to be of service to the schools of the state in other ways.

The results of original investigations carried on by the Bureau of Educational Research are published in the form of bulletins. A list of available publications is given on the back cover of this bulletin. At the present time five or six original investigations are reported each year. The accounts of research conducted elsewhere and other communications to the school men of the state are published in the form of educational research circulars. From ten to fifteen of these are issued each year.

The Bureau is a department of the College of Education. Its immediate direction is vested in a Director, who is also an instructor in the College of Education. Under his supervision research is carried on by other members of the Bureau staff and also by graduates who are working on theses. From this point of view the Bureau of Educational Research is a research laboratory for the College of Education.

BUREAU OF EDUCATIONAL RESEARCH
College of Education
University of Illinois, Urbana

BULLETIN NO. 46

BUREAU OF EDUCATIONAL RESEARCH
COLLEGE OF EDUCATION

THE USE OF SCALES FOR RATING PUPILS' ANSWERS TO THOUGHT QUESTIONS

By

C. W. ODELL

Assistant Director, Bureau of Educational Research

PUBLISHED BY THE UNIVERSITY OF ILLINOIS, URBANA
1929

UNIVERSITY
OF ILLINOIS
PRESS

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

INTRODUCTION

The development of scales for rating pupils' performances. From almost the very beginning of what is commonly called the standardized-test movement, scales as distinguished from tests[1] began to appear. The first test commonly recognized as standardized in the usual sense of the term, was Stone's Arithmetic Reasoning Test,[2] which appeared in 1908. Only slightly more than a year later, Thorndike presented his handwriting scale before Section L of the American Association for the Advancement of Science, and published it within about three months.[3] Two years later, in 1912, Ayres published the first of his handwriting scales,[4] and Hillegas his composition scale.[5] A beginning in drawing was made the next year when Thorndike's scale for measuring achievement in that subject appeared.[6] These four well-known scales were the first of a large number that have been prepared for use in handwriting, English composition, drawing and other subjects.

The problem of this study. This study owes its inception to a suggestion made to the writer that it might be worth while to investigate the possibility and desirability of constructing and using scales for rating pupils' answers to questions in other school subjects than those named in the last paragraph. Such scales would naturally resemble those in English composition more than those in handwriting and drawing and would deal with somewhat similar pupil responses—that is—with thought rather than mere memory answers. As will be shown by data given near the end of Chapter III, experience has revealed that in most cases the use of scales for rating pupils' compositions results in increased reliability of marking. It seemed not unlikely that the same result would ensue if the same technic were applied to other pupil responses more or less similar to compositions. The writer, therefore, decided to undertake the construction and trying out of a number of scales of this sort with a view to ascertaining whether or not the results suggested above would be attained. The general

---

[1]When the words "scale" and "test" are used in contradistinction to each other, the former is ordinarily employed to denote a set of samples or specimens arranged in order of merit with which pupils' performances are to be compared. "Test," on the other hand, refers to a measuring instrument or portion thereof which secures pupils' performances. Thus a series of problems in arithmetic or of questions in history, for example, is a test because it elicits responses from the pupils, whereas a series of specimens of handwriting ranging from very poor to very good, is a scale.

[2]Stone, C. W. "Arithmetical Abilities and Some Factors Determining Them," *Teachers College, Columbia University Contributions to Education,* No. 19. New York: Bureau of Publications, Teachers College, Columbia University, 1908. 101 p.

[3]Thorndike, E. L. "Handwriting," *Teachers College Record,* 11:1-93, March, 1910.

[4]Ayres, L. P. "Scale for Measuring the Quality of Handwriting of School Children," *Russell Sage Foundation, Bulletin* E-113. New York City: Russell Sage Foundation, 1912. 16 p.

[5]Hillegas, M. B. "A Scale for the Measurement of Quality in English Composition by Young People," *Teachers College Record,* 13:331-84, September, 1912.

[6]Thorndike, E. L. "The Measurement of Achievement in Drawing," *Teachers College Record,* 14:345-83, November, 1913.

question to which an answer was sought may be stated as follows: Does the use of scales for rating pupils' answers result in greater reliability of the marks given such responses than if scales were not employed?

In seeking an answer to this question, it was necessary to narrow it and to establish certain limitations. It was, of course, impossible to deal with all types of questions in all subjects, so that the conclusions arrived at apply specifically only to certain scales described in Chapter II which deal with a number of types of questions in several high-school subjects.[7] It was the writer's intention, however, to construct scales that would be as typical as possible, and to deal with such a wide variety of questions and enough different subjects that the results might reasonably be considered as generally applicable.

Plan of this bulletin. Chapter II will be devoted to a rather detailed account of the construction of the scales and the carrying on of the investigation. Chapter III will present the statistical results from the standpoint of comparing ratings with the scales to ratings without them, will attempt to analyze these data, and finally will make a brief comparison with similar data reported by others for English composition scales. Chapter IV will contain a brief summary and the general conclusions drawn from the investigation. In Appendix A, the complete list of questions used in the scales will appear. Appendix B will present data from ratings of sets of pupils' answers rather than from single answers and will discuss briefly the reliability of traditional examinations.

[7] The types of questions and the high-school subjects dealt with will be found in full on p. 8.

CHAPTER II

THE CONDUCT OF THE INVESTIGATION

Purpose of this chapter. It is the purpose of this chapter to give a somewhat detailed account of the construction of the scales and the experimental work carried on with them. The various steps involved in conducting the investigation may be stated as follows:

1. Determining types of questions to be dealt with
2. Selecting questions to be tried out
3. Securing pupils' answers to these questions
4. Preliminary rating of answers
5. Selecting questions for scales
6. Final rating of answers to questions used in scales
7. Selecting answers to be included in scales
8. Preparing criticisms of answers
9. Experimental rating without and with scales

These steps will be taken up in order in the following paragraphs.

1. Determining types of questions to be dealt with. At the very beginning of the investigation it was decided to deal only with what are ordinarily called thought questions. These may be defined as questions that can not be satisfactorily answered by pupils through mere memorization and recall of information, but which require an element of reasoning or reflective thinking to produce satisfactory answers.[1] The next step was to determine what types or kinds of thought questions should be dealt with. As a preliminary step to making this determination, it was necessary to have an analysis or classification of thought questions into the various types thereof. Instead of attempting to make an original classification for this purpose, the writer chose what seemed to him to be, at least for the purposes of this investigation, the most satisfactory one already made. It was that of Monroe and Carter,[2] who, in making a study of the various kinds of thought questions employed in secondary schools, listed and defined twenty different types. These types are as follows:

1. Selective recall
2. Evaluating recall
3. Comparison of two things—on a single designated basis
4. Comparison of two things—in general
5. Decision
6. Cause or effect
7. Explanation
8. Summary
9. Analysis

[1]Although memorization and recall alone do not enable pupils to give satisfactory answers to thought questions, they are ordinarily necessary elements in the ability to answer them. In other words, pupils cannot think unless they remember some facts to use in their thinking.

[2]Monroe, W. S. and Carter, R. E. "The Use of Different Types of Thought Questions in Secondary Schools and Their Relative Difficulty for Students," *University of Illinois Bulletin*, Vol. 20, No. 34, Bureau of Educational Research Bulletin No. 14. Urbana: University of Illinois, 1923. 26 p.

10. Statement of relationships
11. Illustrations or examples
12. Classification
13. Application
14. Discussion
15. Statement of aim
16. Criticism
17. Outline
18. Reorganization of facts
19. Formulation of new questions
20. New methods of procedure

Monroe and Carter did not claim that these twenty types included all possible thought questions, but, for the purposes of the present study, no great importance attaches to this fact.

Inasmuch as the use of as many as twenty types of questions would either have resulted in demanding greater expenditures of time and money than it was felt were justified, or have necessitated dealing with each type of question less thoroughly than was desired, it was decided to select a limited number of the twenty types for use in the investigation. A careful study of the kinds of answers that pupils would probably give to each of the types of questions named by Monroe and Carter resulted in the selection of the following nine of the twenty types for use in the experiment: analysis, cause or effect, comparison, criticism, discussion, explanation, relationship, reorganization, and summary. The answers to these nine types of questions appear to be such as lend themselves more readily to rating by the use of scales than do those to the other eleven types.[3]

2. Selecting questions to be tried out. After deciding upon the types of questions to be used, the next step was to secure a sufficient number of satisfactory questions of each type. Before this could be done, it was necessary to decide upon the school subjects to be dealt with. Partly because of a desire to include rather different types of subject-matter, partly because of the interests of those carrying on the experiment, and partly because of the greater supply of questions readily available in certain subjects than others, it was decided that questions should be prepared in each of four high-school subjects: civics, general science, American history, and English literature. In order to have a large enough number of questions of each type in each subject to render probable that at least one satisfactory question could be chosen from among them, it was decided to prepare a total of two hundred questions, of which there were five or six of each of the nine types in each subject. There happened to be available at the Bureau of Educational Research a rather large collection of examination questions prepared by high-school teachers. Accordingly,

[3]It cannot be stated as a fact that it is easier to rate answers to these nine types of questions by means of scales than it would be to do the same for the eleven other types, but merely that after careful consideration it was the opinion of the writer and his assistant, Mr. J. A. Blough, that this was true.

those in the four subjects named above were carefully examined for questions that seemed suitable for the purposes of the experiment. Some of the questions found were used just as they were, others were more or less modified. The examination of the lists, however, did not produce sufficient numbers of questions for all of the different types and subjects. Two or three of the teachers at the University High School, therefore, were asked to prepare a few questions of the kinds needed and a number of these were included. Also a few were formulated by the writer and his assistant. A list of two hundred questions, fifty in each subject, as nearly as possible equally divided among the nine types, was compiled from these various sources.

3. Securing pupils' answers to these questions. As the first step toward securing pupils' answers to the two hundred questions, those in each subject were divided into five lists, known as lists A, B, C, D, and E, of ten questions each. Each list contained one question of each of the nine types, with an additional question in some one of these types. Letters were then sent to the principals of slightly more than one hundred selected high schools in the state of Illinois asking for their cooperation in the project. Enclosed with each letter was a copy of one of the lists of ten questions for each teacher of each of the four subjects. The principals were requested to ask these teachers to have their pupils write answers to some or all of the questions and to return the answers to the writer. The total number of teachers of the four subjects in the high schools to which the material was sent was about seven hundred fifty, of whom slightly more than three hundred were literature teachers, about two hundred were history teachers, and the remainder almost equally divided between civics and general science. Most of the principals addressed cooperated to the extent of handing the lists of questions to the proper teachers, and many of them also strongly urged the teachers to secure and send in pupils' answers thereto. No directions were sent regarding the conditions under which the questions were to be answered, since for the purposes of the study it was not important that the answers be accurate measures of the pupils' achievement. What was desired was to secure a large number of pupils' answers, varying in merit from none at all up to perfect, for each of the questions upon the lists. The total number of usable answers to the whole 200 questions sent in to the writer was about 23,600, of which approximately 5,400 were in civics, 6,900 in general science, 4,000 in history, and 7,300 in literature. These were rather well distributed among the different types of questions so that for every type in every subject there was at least one question, and usually two or three, to which the number of answers was well over 100.

4. Preliminary rating of answers. As has already been stated, it was desired to secure a rather large number of pupils' answers, ranging

all the way from those of no merit at all to perfect ones, so that a number of specimens ranging at equal, or as nearly equal as possible, intervals from zero to perfect, could be selected for inclusion in the final scales. As the first step, a preliminary rating of each answer was made by a single person. Those who did this rating were all experienced teachers of the subjects dealt with, and two of the four had had some graduate work. All of the rating in each of the four subjects was done by a single person. The plan of rating was to classify the answers to each question into eleven degrees of merit, ranging from zero up to ten. The raters were instructed to pay no attention to handwriting nor to mistakes in English unless they were sufficiently serious to obscure the meaning.

5. Selecting questions for scales. When these preliminary ratings were tabulated, it was found that some of the questions were evidently too difficult, since nearly all of their answers received low ratings, whereas others were too easy, as most of their answers received decidedly high ratings. It had been hoped that after the preliminary rating there would be at least one question of each type in each subject for which nine or more answers had been rated as belonging in each of the eleven degrees of merit from zero to ten, inclusive. In many instances, this expectation was satisfactorily fulfilled, but in others no one of the five or six questions provided a satisfactory distribution of answers. In a number of cases in which the desired conditions were not entirely fulfilled, it was possible to find questions which came so near to having nine answers in each of the eleven degrees of merit, that it seemed satisfactory to include them. In others, however, there were none that approached the desired conditions nearly enough to justify their inclusion. In these cases, however, there were questions which had the desired numbers and distributions of answers except that there were too few at the upper degrees of merit. In order to remedy this situation, letters were addressed to more than three hundred teachers of the four subjects, asking them to prepare and send to the writer what they considered satisfactory answers to some one or two questions which were sent them. The answers prepared and returned by these teachers were then given single ratings and included in the total distributions of answers, thus making enough answers of high degrees of merit. The final selection of the questions to be included in the scales was then made. It was based entirely on the answers finally available, from which those to be included in the scales could be selected, and not directly on the questions themselves, except in so far as this was reflected in the answers received. Thus thirty-six questions,[4] one of each of the nine types in each of the four subjects, were chosen.

[4]The thirty-six questions used in the scales may be found in Appendix A.

6. Final ratings of answers to questions used in scales. As a preliminary step to the final ratings for purposes of determining the exact answers to be included in the scales, the number of answers to each of the thirty-six selected questions was reduced to one hundred. Wherever possible, this number consisted of nine answers at each of the eleven degrees of merit and one additional one taken at random. These thirty-six sets of one hundred answers each were then rated by experienced teachers. All of the English literature answers were rated by five persons, but in the case of the other three subjects the answers were rated by only four, or even three, persons. No rater knew what score had been assigned any particular answer at an earlier rating, nor were the papers arranged in any fixed order with regard to merit.

7. Selecting answers to be included in scales. It had been hoped that, as a result of this final rating, there would be found at least one answer at each of the eleven points on the scale for each of the nine types of questions in each of the four subjects, concerning the merit of which there was unanimous agreement. That is, it had been hoped that all of the raters would give the same rating to at least one paper at each value. This result was not achieved, however, in even a majority of instances. In many, there was no one of the nine papers given the same preliminary rating concerning which this was true. In case there was only one on which the opinions were unanimous, this was selected as the one to be included in the scale. If there was unanimous agreement concerning more than one, ordinarily the shorter one of the two was chosen, although this was not always true. In case there was no answer concerning which all the raters were agreed, the one was selected upon which they were most nearly agreed. Whenever possible, this was an answer upon which three of the judges, in cases where there were five, had given the same rating, one of the other two a rating one point above, and the fifth a rating one point below. For example, three of the judges might have rated a specimen as of merit 6, one of 7, and one of 5.

In order that the reader may see about how great was the degree of agreement among the raters who gave the final ratings, a typical set of such ratings for the eleven answers in one scale is given in Table I. This shows, for example, that the first four of the raters gave the first answer 10, but Rater E gave it only 9. For the answer which was placed at 9 on the scale there were three ratings of 9, one of 10, and one of 8. In the case of only three of the eleven answers, those placed at 7, 5, and 0, was there unanimous agreement.

One slight exception to the procedure described above should be noted. In a very few cases, there were no answers which half or more of the raters agreed deserved a rating of ten or of zero. In

TABLE I. FINAL RATINGS GIVEN THE ELEVEN ANSWERS IN ONE SCALE BY FIVE RATERS

Value at which placed	Raters				
	A	B	C	D	E
10	10	10	10	10	9
9	10	9	9	8	9
8	8	8	8	8	7
7	7	7	7	7	7
6	7	6	7	5	6
5	5	5	5	5	5
4	5	5	4	4	2
3	3	4	2	3	3
2	3	2	2	2	1
1	3	1	0	1	1
0	0	0	0	0	0

cases of the first sort, those who did not think the specimen worth ten were asked to state their reasons, and whatever changes were necessary to make them give ratings of ten were made. Similarly, with specimens at the lower end of the scale, the raters who placed them above zero were asked to state the points upon which they allowed credit and these were either eliminated or modified so as to be incorrect.

8. Preparing criticisms of answers. After it had been decided which answers should be included at each of the eleven degrees of merit for each scale, the next step was the preparation of brief criticisms upon the answers. These criticisms were intended to point out why each answer was better than those below it on the same scale and not so good as those above it, and thus to aid persons employing the scales in distinguishing between different degrees of merit. Each of the persons who had taken part in the final rating was asked to compare the answers and prepare a set of such criticisms. These were then assembled and the several criticisms of each answer combined into a harmonious whole.[5] With the completion of these criticisms, the scales were ready for publication and were sent to the printer.[6] They were published in the form of 8½ by 11 inch pamphlets, one for each of the four subjects. All of a single scale—that is, the question, the answers at each value from zero to ten, inclusive, and the criticisms of these answers—were printed on a single double page so that all of a single scale could be easily visible at once. To show more clearly what the scales are like, a complete copy of that for reorganization in literature is given.

[5]The combination and final preparation of criticisms was done by the writer with the help of Mr. E. H. Sanguinet, at that time his assistant.

[6]In reproducing the pupils' answers contained in the scales, all errors in grammar, spelling, and so forth, were retained so that they would be as nearly as possible typical of answers actually obtained in ordinary school work.

TYPE VIII.—REORGANIZATION

Show in order the changes that came over Sir Launfal in his search for the Holy Grail

Value 0

Pupil Answer. He was dream all of this time that he thought he was hunting for the holy grail and about running on to this old man that had found it.

Criticism. This appears to be an attempt to give a brief summary of the plot rather than to show the changes that came over Sir Launfal. No such change is shown nor is any trait of character mentioned, neither is any conception of what the search for the Holy Grail was indicated.

Value 1

Pupil Answer. Sir Launfall in his search for the Holy Grail changed after he saw and helped the leper. He saw that there was good in everything and that the size or value of anything meant nothing to God.

Criticism. This mentions that after seeing and helping the leper Sir Launfal saw the good in everything. Apparently the writer had some idea of the changes that came into his life and some understanding of his search for the Holy Grail. Sir Launfal's early attitude of pride and his flinging the coin to the beggar are omitted. It is not plainly stated how Sir Launfal changed, the last sentence not being clear.

Value 2

Pupil Answer. When Sir Launfal first started out for his search for the Holy Grail he was hard hearted and flung to the beggar a coin, but when the beggar told him the coin was not what he wanted that it was something to eat, Sir Launfal saw and understood and became a changed man.

Criticism. This describes Sir Launfal's early attitude and his flinging the coin to the hungry beggar and thus shows his character at the beginning more definitely. Its chief merit is that it is longer and gives more details. Sir Launfal's struggle with hardships and the results are not mentioned. A false statement is made to the effect that the beggar told him he did not want the coin, but something to eat. It is implied that meeting the leper was all that was needed to cause the change in Sir Launfal.

Value 3

Pupil Answer. When Sir Launfal set out on his search for the Holy Grail he was very proud and hated the beggar that sat by the Gate when he returned he had gone through many hardships himself and his soul was filled with love for others.

Criticism. This speaks of the subjugation of Sir Launfal's pride through the hardships he endured. It is more definite in giving the cause, results and order of changes than is Specimen 2. It is not stated that Sir Launfal found the Grail when he shared his crust with the beggar. The idea of hatred is brought in where it does not belong. The search for the Holy Grail is mentioned, but practically no details are given.

Value 4

Pupil Answer. When Sir Launfal started out he scorned the beggar at the gate. He was dressed in a very rich suit of armor. After he had become an old man in the search he didnot scorn the beggar any more but shared his bread and water. It was then that he found the Holy Grail.

Criticism. The writer brings out the contrast between Sir Launfal when young, rich and proud and later when old, poor and sympathetic. The statement of his search for the Holy Grail is fairly definite. The sharing of the bread is said to be symbolic of the change in his character. It is not brought out that his early failure to find the Grail was due to his thinking himself above the leper. The cause and manner of change are both treated indefinitely.

Value 5

Pupil Answer. Sir Lanful went out in search of the Holy Grail, he was a young man in and very proud, too proud to look at the leper that crouched at his gate, but threw him a coin. On returning from his search, tired and worn out, he saw the leper still there and learned that the Holy Grail was at his own door but he had passed it by in thinking himself "above" the leper.

Criticism. This answer shows that Sir Launfal's early failure to find the Grail was because he thought himself above the leper. His early pride is well brought out and the reason for the change is stated more definitely than in most of the poorer answers. It is not, however, shown that the years of discouraging search rendered him sympathetic enough to share his crust with the beggar. Although it is stated that he finally discovered the Holy Grail at home, it is not indicated how he did so. On the whole the description of him is very poor.

VALUE 6

Pupil Answer. At first Sir Launfal was a dignified man of wonderful appearance, and very wealthy. Next we see him pass the beggar at the gate and give him money in scorn. He searches for the Holy Grail, and after years of weary journeys he becomes sympathetic. Next he returns home a ragged worn out beggar, and now confronts the beggar at the gate of the castle, once his, now lost, and he shares with him his bread.

Criticism. The search for the Holy Grail is explained and the change from pride to selfishness brought out. The reasons for the change are more definitely given than in Specimen 5. Likewise this makes clear just how the Holy Grail was found. It does not, however, draw the distinction between giving alms from duty and from a real desire to help, nor does it show that it was Sir Launfal's learning the real meaning of love that made him richer when he returned than when he started out. There is a tendency to tell events rather than to show changes.

VALUE 7

Pupil Answer. It was a young knight, clad in bright knights armor that Sir Launfal started on his search for the Holy Grail He knew the rules of knighthood but his proud heart did not know the joy of practicing them. For instance when he gave to the beggar it was not from the heart but merely to get it done that he might go on his way.

Many times during his long fruitless quest he met with people in need of something which he might have given with love but it was not until years later when he returned, a poor old man, without the Holy Grail that he learned the meaning of love. Now a beggar, himself, he was richer in his heart of gold than he had been when he started out in his shining armor.

Criticism. This includes a good description of Sir Launfal when young and rich and of the later changes in him. It brings out the difference between doing good deeds from a sense of duty and from a love of humanity, the last sentence especially doing this well. The accompanying changes are more or less indicated, but the final change in him is not completely described, his later humility and welcome to all being omitted. Too much of the discussion is devoted to his condition before he began to change.

VALUE 8

Pupil Answer. Sir Launfal start in the summer in search of the Holy Grail. He is a young man dressed in beautiful armor. He meets a Leper at his gate. He throws him a piece of gold in scorn. This shows his proud spirit and selfishness. After many years of hunting he returns an old man bent and gray. It is winter. He is turned away from his own door. The leper is at his side. He has a piece of crumb. He breaks the crust and shares it with the leper. He breaks the ice in the stream and gives the leper water to drink with a wooden bowl. He is humble now. The bowl turns into the Holy Grail. Sir Launfal is a changed man. He welcomes all to his door now.

Criticism. This brings out Sir Launfal's later humility and friendship for humanity, but does not give definite facts as to the manner, extent and order of the changes. Definite examples which show the changes are given. Less significance is attached to the character changes than in Specimens 9 and 10. Although the steps in his search for the Holy Grail are given, the accompanying changes that come over him are not.

VALUE 9

Pupil Answer. Sir Launfal's fruitless search and consequent disappointment doubtless made him less proud and less sure of himself than he was when he scorned the leper at the castle gate. Accordingly when he saw the leper on the desert he was touched by the poor man's distress. Brooding perhaps over the cause of his failure he had come nearer to the truth in regard to it, than he realized until the meeting with the leper called for an outward expression of this inward revelation. When he acknowledged that in this man, he recognized the "image of Him who died on tree," he showed that the lesson of humanity had been learned. With the exchange of glances between Sir Launfal and the leper, the knight felt humiliated as he thought of his former arrogance. The kind voice of the glorified leper calmed his restless troubled soul, and soon after he awakened from his slumber, he entered the castle hall, to demonstrate by his kindness and generosity that he had in reality found the Holy Grail.

Criticism. The later changes in Sir Launfal are shown more clearly than in Specimen 8 and a better conclusion is given. This contains rather good character analysis and shows fairly well the significance of events as they relate to the change. Some import-

ant points such as the sharing of the bread and water and the fact that Sir Launfal learned the lesson of the brotherhood of man are omitted. There is also a failure to picture Sir Launfal before the change, thus not making clear just what the change was."

VALUE 10

Pupil Answer. In the beginning Sir Launfal was young, strong, handsome, happy, and richly dressed. He was a seeker for worldly glory. He tossed an alms to the beggar from a sense of duty, but he recoiled from human wretchedness as typified by the poor leper. He was self-satisfied. He felt that the Grail could not be found at home, but must be sought in distant lands.

Sir Launfal returned an old, bent, worn, and frail old man, poor in pocket and discouraged in spirit. He met the beggar, but now, with his heart filled with love, he shared with the leper, whom he recognized as a man and a brother. He was humble instead of self-satisfied. Sharing his bread and water, he found the Grail. There was no need now of going to distant lands, but, waking from his dream, Sir Launfal ordered his armor put away, while he stayed to share his castle with the poor. Again he was happy, though in a joyous peaceful spirit instead of his former haughtiness.

Criticism. The writer brings out the important changes in Sir Launfal and shows that he had learned the lesson of the brotherhood of man, thereby finding the Holy Grail at his own door. Both the reasons for the change and its effects are given. Especially in the second paragraph is the change made vivid and described with appropriate adjectives. It is also made clear that the story is a dream from which he awakens.

9. Experimental rating without and with scales. Most of the experimental work having to do with the determination of whether or not the use of such scales makes marking more reliable was carried on at the offices of the Bureau of Educational Research. A number of teachers of varying degrees of experience were engaged in this work during a period of more than a year. It was planned that each of these persons should rate a number of papers without using the scales, then after an interval of at least a month or two had elapsed, so that there would be little or no danger of recalling the scores given, rate them again without using the scales, then after a second interval rate them with the scales, and finally, after still another interval, a second time with the scales. The papers employed were in sets of twenty-five each, chosen at random from among the pupils' answers to some one question of each type in each subject, but not to the same questions as had been used in the scales themselves. It was intended that each rater should rate one set of twenty-five papers for each of the nine types in a particular subject four times, as outlined above. Some of the raters not only did this for one subject but for more, one of them even doing it for all four. One or two of the raters did not complete their work even on one subject.

In addition to the ratings given by these few people working under close supervision, a number were secured from students enrolled in the course in Technic of Teaching at the University of Illinois. These students were almost all juniors and seniors who intended to enter high-school teaching. Only a few had already had teaching experience, and most of these had taught only a year or two. Since it was not practicable to have these students carry through the complete rating program, each one merely rated one set of twenty-five answers twice, first without the scales, and some time later with them. Thus

the data from these students gave no information as to reliability of ratings by the same student without and with the scales, but merely of the agreement between the ratings given by different students under the two kinds of conditions.

It had been hoped that a considerable number of teachers in actual service would participate in this part of the study, but this expectation was not fulfilled. Although letters containing appeals for such cooperation were sent out along with copies of the scales to all teachers of the four subjects in public high schools in the state of Illinois, the response received in terms of actual rating done was practically nil. A number of responses to another request in the same letter for criticisms of the scales came to the writer, but only one teacher contributed usable data on reliability.

The directions given those working at the Bureau of Educational Research and also the students who did the rating were practically the same as had been given to those doing the preliminary rating.[7] For the ratings with the scales, additional directions similar to those ordinarily given in connection with the use of English composition, handwriting, and other such scales were given. The raters were instructed to compare each pupil's answer with the specimens contained on the appropriate scale and to give it the value of the one that it most resembled in merit as an answer to the question asked. It was suggested that if they were very doubtful it would be well to begin by comparing the answer to be rated with each specimen on the scale from the bottom up until what seemed to be the proper value was reached, and then to do the same from the top down, thus arriving at what seemed to be the best rating.

The total number of ratings of pupils' answers in sets of twenty-five contributed by all persons doing any rating was about 13,500. These ratings were given a total of 1,050 answers, almost all of which were rated once with and once without the scales, and most of which were rated twice with and twice without the scales by from three to six different persons. A total of fifty-seven individuals participated in the rating, six of whom worked in the offices of the Bureau of Educational Research and contributed slightly over 80 per cent of the ratings. Three of these six had also participated in the work done in the construction of the scales—that is, in the rating of the answers from which those included in the scales were selected—whereas the other three had had no previous contact with the scales, and in fact were purposely kept from becoming familiar with them until after they had made their ratings without the scales.

In addition to the data on reliability resulting from the ratings just described, all of which were made after the completion of the scales, the ratings given when the selected sets of one hundred answers

[7]These directions will be found on p. 10.

each were being rated by several different individuals[8] were available. All these ratings were, of course, given without the use of the scales, since they were made prior to their construction. Therefore, they in themselves do not furnish data by which ratings without and with the scales may be compared. To secure such data, a number of sets of one hundred each were selected at random and each re-rated by one or more of the persons who had rated the papers previously. The total number of these ratings was fifty-four hundred, half without and half with the scales.

Summary. This chapter has been devoted to an account of the construction of the scales and the experimental work carried on to determine whether or not their use results in increasing the reliability of ratings given pupils' answers. Nine of the twenty types of thought questions listed by Monroe and Carter[9] were selected as calling for answers which would most readily lend themselves to being rated by scales. Five or six questions of each of the nine types in each of four high-school subjects were prepared and distributed to more than one hundred different high schools. Out of the total of 23,500 pupils' answers, thirty-six sets of one hundred answers—that is, a set of answers to one question of each type in each subject—were selected by a preliminary rating. These sets were rated further by several persons, and finally eleven answers were selected out of each set. Each of these answers had been rated as most nearly possessing one of the eleven degrees of merit—from zero to ten, inclusive. Appropriate criticisms of these answers were prepared, and each set of eleven answers with the criticisms concerning them were combined to form a scale. Almost sixty different persons then rated pupils' answers to questions not actually included in the scales but similar to the ones used therein, both without and with the scales, in order to determine which method of procedure was more reliable. Also ratings given by six persons to answers to questions in the scales were secured and compiled.

[8]For an account of this rating see p. 11.
[9]*See* p. 7.

CHAPTER III

THE RESULTS OF THE INVESTIGATION

The general results. Since the carrying on of the investigation has been described in some detail in the previous chapter, no account thereof will be given here, but instead the writer will proceed at once to present the results. Contrary to what is probably the most common order, the general results—that is, the combined or average results for the whole study—will be presented first, and later these will be analyzed on several different bases.

Three measures of the reliability of ratings were computed. These were the coefficient of correlation[1] between series of ratings given the same answers, the mean differences[2] between ratings, and the differences in mean scores[3] given sets of pupils' answers. Each of these measures reliability in a somewhat different way. To understand the difference, one must be familiar with the distinction between variable errors and constant errors. Variable errors, sometimes called accidental or chance errors, differ for the individuals composing a group. The variable error in one pupil's score may be large, in that of another pupil's, small, in one case positive, in another case negative, and so on. Constant errors, on the other hand, are those that tend to be the same for all members of a particular group. They result from some common cause affecting the performance of all the pupils in the group. The coefficient of correlation is not affected by constant errors and is, therefore, a measure of variable errors alone. The difference in mean scores, on the contrary, is a measure of constant errors. In this study it shows how much more severely or leniently one rater marked than another. The mean difference is affected by both variable and constant errors and thus in a sense offers the same evidence as the coefficient of correlation and the difference in mean scores combined.

The average coefficient of correlation for the 13,375 ratings of the sets of twenty-five answers each was .87 without the scales and .89

---

[1]The coefficient of correlation is an index of rectilinear or straight-line correlation or relationship between two series of paired facts. It ranges in value from + 1.00, which indicates perfect positive correlation, through zero, which indicates no correlation at all, to — 1.00, which indicates perfect negative correlation. For a more complete discussion of the interpretation of the coefficient of correlation *see*:

Odell, C. W. "The Interpretation of the Probable Error and the Coefficient of Correlation," *University of Illinois Bulletin*, Vol. 23, No. 52, Bureau of Educational Research Bulletin No. 32. Urbana: University of Illinois, 1926. 49 p.

The method of computing it may be found in any standard text on statistics.

[2]The term "mean difference" is used to refer to the mean or average difference between two series of ratings given the same answers. It is found by taking the sum of the differences between the two ratings given each answer and dividing by the number of answers concerned.

[3]The "differences in mean scores" are the differences between the mean or average scores given whole sets of answers at two or more different ratings. Thus a difference in mean scores can never be greater than the mean difference for the same answers. In most cases it is smaller, since usually some of the individual differences are positive and others negative, and thus they to some extent balance one another in their effect upon the mean difference.

TABLE II. AVERAGE MEASURES OF RELIABILITY OF RATING PUPILS' ANSWERS WITHOUT AND WITH SCALES BY THE SAME AND BY DIFFERENT RATERS

Raters	Coefficient of Correlation[a]		Mean Difference		Difference in Mean Scores	
	Without	With	Without	With	Without	With
Same	.90	.92	.79	.73	.33	.31
Different	.74	.75	1.86	1.68	1.08	.90

[a]The term "coefficient of reliability" might have been used instead of coefficient of correlation. In its most limited sense, however, coefficient of reliability should be used only to refer to correlation between two or more sets of ratings given by the same persons and not to those given by different persons. For this reason it was not used as a heading.

with the scales.[4] The corresponding mean differences were 1.06 and .94, respectively, and the differences in mean scores .48 and .43. Since a larger coefficient of correlation and a smaller mean difference or difference in mean scores indicates greater reliability, the figures just given show that on the whole the ratings with the scales were slightly more reliable than those without the scales. The difference in reliability was, however, too small to be significant.[5] It certainly cannot be said that an increase in the coefficient of correlation of only .02 or decreases as small as those found in the other measures indicate that the use of the scales employed in this investigation increases reliability of marking enough to be justified. Since, however, it is possible that in combining the results for all the ratings of the sets of twenty-five, important factors have been lost sight of, the data will be grouped and analyzed on different bases and thus considered further in the following paragraphs.

Ratings by the same and by different raters. One possibility is that if repeated ratings by the same persons are separated from those by different raters, some difference may be found; therefore, this separation was made. The results are presented in Table II. This shows, for example, that the average coefficient of correlation of the ratings of the same raters was .90 without the scales, and .92 with them. From all the figures given, it appears that the results from the same scorers and those from different scorers constitute similar evidence as to the comparative reliability of ratings without and with the

[4]The data reported in this chapter were computed from ratings of answers to single questions. Data from combined ratings of answers to a number of questions such as would compose a single traditional examination will be found in Appendix B. It will be seen that the conclusions toward which they point are the same as those drawn from the ratings of answers to single questions.

[5]In the case of practically all coefficients of correlation, mean differences, and differences in mean scores given in this bulletin, the numbers of cases upon which they are based are decidedly large. Therefore, the probable errors of these measures are very small, in most cases being less than .01 for coefficients of correlation and correspondingly small for the other measures. Since they are so uniformly small, they will not be given from time to time.

TABLE III. AVERAGE MEASURES OF RELIABILITY OF RATING PUPILS' ANSWERS WITHOUT AND WITH SCALES FOR THE NINE TYPES OF QUESTIONS

Type of Question	Coefficient of Correlation		Mean Difference		Difference in Mean Scores	
	Without	With	Without	With	Without	With
Analysis	.85	.89	1.01	.86	.59	.35
Cause or Effect	.84	.88	1.52	.90	.51	.44
Comparison	.86	.85	1.05	1.23	.49	.50
Criticism	.86	.91	1.07	.86	.58	.37
Discussion	.86	.88	1.05	.90	.40	.40
Explanation	.90	.89	.99	1.03	.51	.60
Relationship	.86	.87	1.00	1.03	.40	.40
Reorganization	.87	.89	.96	.92	.49	.42
Summary	.90	.93	.88	.72	.40	.35
Average	.87	.89	1.06	.94	.48	.43

TABLE IV. AVERAGE MEASURES OF RELIABILITY OF RATING PUPILS' ANSWERS WITHOUT AND WITH SCALES IN THE FOUR SUBJECTS

Subject	Coefficient of Correlation		Mean Difference		Difference in Mean Scores	
	Without	With	Without	With	Without	With
Civics	.88	.91	.84	.80	.30	.40
General Science	.87	.89	1.05	.91	.49	.40
History	.84	.86	1.03	.98	.95	.51
Literature	.87	.90	1.10	.94	.64	.41
Average	.87	.89	1.06	.94	.48	.43

scales and agree with the combined results. As would be expected, the reliability of ratings by the same scorers is decidedly higher than that of those by different scorers, but in each case there is a slight increase in the reliability of ratings with the scales over those given without the use of the scales. Thus a separation of the data on this basis does not result in any conclusions additional to those drawn from the combined data.

Ratings of questions of different types and in different subjects. Another possible basis of grouping the data is according to the nine types of questions and the four subjects dealt with. The average measures of reliability for the different types of questions are given in Table III and those for the four subjects in Table IV, which are similar in general form to Table II.

An inspection of the first of these two tables makes it evident that there are no great differences among the various types of ques-

Table V. Average Measures of Reliability of Rating Pupils' Answers Without and With Scales by Experienced Teachers and by Those Without Teaching Experience

Raters	Coefficient of Correlation		Mean Difference		Difference in Mean Scores	
	Without	With	Without	With	Without	With
With Experience.....	.79	.80	1.48	1.47	.74	.74
Without Experience.....	.68	.69	2.17	1.81	1.72	.91

tions. For two or three types, ratings without the scales appear to be fully as reliable as those with the scales, or even slightly more reliable; whereas, in the case of several other types, the differences in favor of the use of the scales are somewhat greater than the average difference. None of the differences are, however, great enough to warrant the conclusion that the type of question dealt with is an important consideration.

Table IV shows that the results in the four subjects are very similar. The data for each indicate that rating with the scales is a very little more reliable than that without it. From those in both tables it appears that when the combined figures for reliability are analyzed according to type of question or subject, nothing is shown in addition to what is revealed by the combined results.

Ratings by experienced teachers and by those without experience. Another possibility is that a difference may be found between ratings by experienced teachers and by persons who had no teaching experience. As was stated in Chapter II, very few of the students in Technic of Teaching had had such experience, and of those who had, none had had any considerable amount. All of the other persons who rated the answers had taught varying lengths of time up to about twenty years. Table V, similar to II, III, and IV, has been constructed to show what differences, if any, in rating without and with the scales existed for these two groups of raters.

The figures in the first row of this table indicate that for the raters with teaching experience there was very little difference in the ratings without and with the scales. For those without teaching experience, the coefficients of correlation show little difference, but the mean differences and the differences in mean scores indicate that ratings with the scales are more reliable by significant amounts. This is especially true of the differences in the mean scores, since the average difference with the scales, .91, is only slightly more than half the average difference without the scales, 1.72. This seems to point to the conclusion that experienced teachers, who presumably have already developed

TABLE VI. AVERAGE MEASURES OF RELIABILITY OF RATING PUPILS' ANSWERS WITHOUT AND WITH SCALES BY PARTICIPANTS AND NON-PARTICIPANTS IN PRELIMINARY RATING

Raters	Coefficient of Correlation		Mean Difference		Difference in Mean Scores	
	Without	With	Without	With	Without	With
Same						
Participants	.89	.93	.74	.57	.30	.32
Non-Participants	.88	.91	.92	.70	.28	.40
Different						
Participants	.79	.78	1.60	1.47	1.08	.72
Non-Participants	.77	.80	1.51	1.48	.64	.74

more or less fixed and satisfactory habits of rating, do not profit to any appreciable degree by the use of the scales in so far as the reliability of their ratings is concerned, but that those without teaching experience profit in that the standards in mind at the times of the two ratings are much more nearly the same when the scales are employed. In other words, the scales appear to have a considerable influence with inexperienced raters in fixing general standards, but not in determining the ratings given individual answers.

Ratings by those who had participated in constructing the scales and by those who had not. Of the six experienced teachers who did the major part of the rating of the sets of twenty-five, three had participated in the rating necessary in the construction of the scales,[6] whereas the other three had not. The writer, therefore, tabulated separately the reliability measures for these two groups of three each without and with the scales to endeavor to ascertain if there were any significant conclusions to be drawn therefrom. These figures are given in Table VI.

In this table, which is similar to Tables II, III, IV, and V with regard to the measures of reliability given, those of participants and non-participants in the rating have been separated according to whether the answers were re-rated without and with the scales by the same scorers or by different scorers. The differences are somewhat irregular, and do not appear to justify any conclusions to the effect that ratings without and with the scales are significantly different for the two groups of raters. Therefore it appears that previous participation in the making of the scales was not an important factor with regard to the reliability of ratings made without and with the scales.

Ratings of answers to the questions in the scales and to questions not in the scales. As was stated in the description of the experimental

[6] *See* p. 11 for an account of this rating.

TABLE VII. AVERAGE MEASURES OF RELIABILITY OF RATING PUPILS' ANSWERS TO QUESTIONS IN THE SCALES WITHOUT AND WITH SCALES

Raters	Coefficient of Correlation		Mean Difference		Difference in Mean Scores	
	Without	With	Without	With	Without	With
Same	.92	.87	.89	1.10	.14	.50
Different	.76	.73	1.51	1.52	.24	.16

rating, the sets of twenty-five, the rating of which furnished the basis of the data so far presented, were answers to questions not in the scales. There is a possibility that if they had been answers to the same questions that were included therein, the raters would have profited more by using the scales in the sense that their ratings would have been more reliable. Ratings made without the scales were evidently not affected, since, in such cases, it was immaterial whether the answers were to questions in the scales or not. In order to ascertain if this factor was productive of different results, re-ratings were given a random selection of a number of the sets of one hundred answers from which those employed on the scales were selected. The reliability measures resulting from these re-ratings are presented in Table VII.

The figures in this table not only support the conclusion indicated by the results from rating the sets of twenty-five, that the use of the scales appears to result in no significant increase in reliability, but even show a decrease therein. For both the same and different raters, the coefficients of correlation are smaller and the mean differences larger with the scales than without them. The differences in mean scores are at variance, in one case that with the scales being greater, in the other case, that without them. Certainly there is no indication at all that the scales are more helpful in increasing reliability of rating if the answers dealt with are to the questions included in the scales than if they are to other questions of the same types.

Ratings by a high-school teacher. As was stated near the end of Chapter II, only one teacher actually in service responded to the request for cooperation with usable data concerning the reliability of rating without and with the scales. This teacher rated several sets of pupils' answers in general science, using the same procedure as was followed by those who made the ratings at the Bureau of Educational Research.[7] The average coefficient of correlation from his ratings without the scales was .68 and from those with the scales .63. The cor-

[7] *See* p. 15 for a description of this procedure.

responding mean differences were .43 and 1.24, and the differences in mean scores .19 and .54. Thus it appears that this teacher's ratings with the scales were less reliable than those without them. The reason for this appears to be that he had a well-planned point system of marking in the use of which he had become relatively expert, and that an attempt to use a new system resulted in disarranging his habitual procedure without substituting anything as good or better in its place.

Reported results from similar studies with English composition scales. Before proceeding to draw any final conclusions, the results obtained in this investigation will be compared with those obtained from similar experiments with English composition scales. As a basis for this comparison, the writer will summarize briefly the reported results from a few such studies.

Ruch and Stoddard[8] report data for a carefully conducted experiment with a limited number (fifty) of pupils as follows: The coefficient of correlation between ratings given by two teachers not employing a scale was .80. Ratings made with various scales yielded correlations from .40 to .92, the average being about .62. In other words, these data seem to show not only no gain in reliability with the use of the scales, but a positive loss.

Among those who have done the most work in this field is Hudelson. In one article[9] he reports, along with other data, figures for 157 judges who rated the same number of themes. Results are given in terms of the average deviation[10] of the ratings from the average or supposedly true values of the themes. The teachers first rated the themes without using a scale and then re-rated them five times with a scale,[11] each time after the acquisition of greater familiarity with it. The average deviation when the themes were rated at first, without a scale, was 1.13. After the scale was merely read, there was a slight increase, the average deviation being 1.22. After it was studied, discussed, and applied for two hours, the average deviation dropped to .77, after four hours to .55, after six hours to .21, and after sixteen hours to .18. These figures indicate that a very definite increase in reliability accompanied greater familiarity with the scale.

Another worker who has reported results of the same sort is Theisen.[12] He had fifteen teachers rate twelve specimens of English composition according to the ordinary percentile system and later with

---

[8]Ruch, G. M. and Stoddard, G. D. *Tests and Measurements in High School Instruction.* Yonkers: World Book Company, 1927, p. 28.

[9]Hudelson, Earl. "The Effect of Objective Standards upon Composition Teachers' Judgments," *Journal of Educational Research,* 12:329-40, December, 1925.

[10]The average deviation, better called the mean deviation, is merely the ordinary arithmetic average of all deviations or differences between the various scores or measures in a distribution and the average. In other words, the deviations or differences of the scores from the average are added and their sum divided by the number of scores.

[11]The scale used in this case was the Hudelson Typical Composition Scale.

[12]Theisen, W. W. "Improving Teachers' Estimates of Composition Specimens with the Aid of the Trabue Nassau County Scale," *School and Society,* 7:143-50, February 2, 1918.

an English composition scale.[13] The variations of the teachers' ratings from the standard value of each specimen[14] were then determined. The average variation of the ratings made without the scales was 19.1 per cent and of those with the scales, 11.6. There were only two of the twelve specimens in which the average variation of ratings with the scale was greater than that of those without the scale and in one of these the difference was very small.

Gordon[15] is another who has reported data bearing on the same point. She had twenty-five compositions rated by forty-one students in educational psychology who had had no previous experience in rating such compositions. Their standard deviation[16] was 11.2. A few days later, the use of a composition scale was explained and the students re-rated the same compositions with the Hillegas Scale. The standard deviation of these ratings was 8.9. On the other hand, when the students were paired and coefficients of correlation computed between their ratings of the compositions, the average coefficient for ratings without the scale was .48 and that for ratings with the scale only .46. They were also divided into two groups of twenty and twenty-one members, respectively, and the average scores of these groups correlated. This yielded coefficients of .87 both without and with the scale.

Other studies containing results that bear upon the point at issue might be referred to, but it is probable that the four which have been briefly summarized are sufficient for the present purpose. Probably most studies of this sort show gains in reliability when scales are used, and some, such as Hudelson's, very marked gains, but others report no gains at all or even losses.

Comparison of results in this investigation with those reported for English composition scales. On the whole, it appears that the results in this investigation are not as favorable to the use of scales as are those reported for English composition. Those in some of the English composition studies agree rather closely with those found by the writer, but in others they indicate that a decided increase in reliability results from employing scales. If the conditions of the various experiments are analyzed, it seems that in a number of instances one rather important differentiating factor may be found. In most cases in which the use of composition scales has resulted in marked increases in reliability, it appears that those who employed the scales

---

[13]The scale employed was the Nassau County Supplement to the Hillegas Scale.

[14]The twelve specimens employed by Theisen were selected from a larger number which Thorndike had arranged for use in such experiments and for which he had determined standard values.

[15]Gordon, Kate. "A Class Experiment with the Hillegas Scale," *Journal of Educational Psychology*, 9:511-13, November, 1918.

[16]The standard deviation is the deviation or distance from the average which includes 34.13 per cent of all cases in the distribution on each side of the average. In other words, slightly more than two-thirds ($2 \times 34.13 = 68.26$) of all the cases are within one standard deviation of the average.

made a more or less careful and well-directed study of them and their use. In other cases in which they merely saw the scales or even had some practice in using them without much study, discussion, or direction, the results were less favorable to the use of scales. In the investigation described in this bulletin, none of the raters had any extensive instruction in the use of the scales nor did they participate in any considerable discussion thereof. It is evident, however, that the three final raters who also participated in the preliminary rating had a very considerable amount of experience in using the scales, and that even those who participated only in the final ratings acquired a fair amount of experience before these ratings were completed. It would appear, therefore, both from this study and from those reported in English composition, that mere practice with scales is not in itself sufficient to insure increased reliability of rating.

One point of difference between English composition scales and those constructed by the writer which may account for more favorable results with the former in some cases should be noted. In the case of most scales in English composition, the values assigned the samples included were determined from the ratings of a much greater number of persons than was the case for the writer's scales. In other words, the values of the specimens in the scales are more reliable, and, therefore, it is more likely that persons employing them will differ somewhat less as to the ratings to be assigned pupils' responses.

Summary. When all the data from the rating of the sets of twenty-five answers are combined, the average reliability measures of ratings made with the scales are only very slightly higher than are those of ratings without the scales. A separation on the basis of ratings by the same and by different persons does not result in any different conclusions, neither does a division according to types of questions and subjects. When the ratings were grouped according to whether those giving them were experienced teachers or not, it was found that for the latter the use of the scales appeared to be helpful in fixing general standards, but not in increasing the reliability of ratings of individual papers. Whether or not the raters had participated in the construction of the scales was not significant. Contrary to what would seem to be probable, ratings of answers to the questions contained in the scales were less reliable when the scales were used than when they were not. A survey of results obtained from similar experiments with English composition scales indicates that in some cases the use of scales does not appear to increase reliability, whereas in others it seems to result in decided increases. Apparently the chief differentiating factor is the definite study of the scales and how to employ them, as distinguished from mere repeated use of them. Such careful study appears usually to result in increased reliability of ratings.

CHAPTER IV

SUMMARY AND CONCLUSIONS

Almost as soon as the standardized-test movement began, scales for measuring pupils' handwriting, drawing, English composition, and similar abilities began to appear. The scale idea, however, has received practically no application to the rating of pupils' answers to ordinary examination questions. In order to ascertain whether or not the use of scales for this purpose would increase the reliability of marking, the writer carried on the experimental work described in this bulletin. Nine types of thought questions were chosen and a number of questions in each of these types were prepared for civics, general science, American history, and English literature. Pupils' answers to these questions were secured and given a preliminary rating. On the basis of this rating, one question of each of the nine types in each of the four subjects was chosen for inclusion in the scales. Eleven answers to each of the questions so chosen were selected, which, according to the average judgment of several raters, most nearly deserved ratings of 0, 1, 2, and so on up to 10. Criticisms of these answers were prepared. Each question, with the set of eleven answers thereto and the accompanying criticisms, formed a scale. Over twenty-three thousand pupil answers to questions similar to but not identical with those in the scales and also over five thousand answers to some of the questions in the scales were then rated both without and with the scales. The results were then compared in order to ascertain whether or not the use of the scales served to increase reliability of rating.

On the whole, the reliability of ratings given with the scales was not found to be significantly higher than that of those given without the scales. This is true not only for the combined data from all of the ratings, but also when these data were separated and analyzed on various different bases. The only real exception to this is that in the case of raters who had no teaching experience it appears that the scales tended to assist in fixing general standards although not to increase the reliability of rating single answers.

A comparison of the results from this investigation with results reported by a number of persons who have made similar studies with English composition scales indicates that in some cases the use of scales for rating English compositions produces a decidedly greater gain in reliability than was found by the writer. Apparently the chief factor producing this difference is that, in the studies which yielded these results, the persons doing the rating had made a rather careful study of the scales in addition to acquiring practice in employing them. It is also possible that better results are obtained with the

English composition scales because the values of the specimens included in them have been determined more accurately than was true for those in the writer's scales. Certainly the writer would not state as a definite conclusion of this study that the use of scales for rating pupils' answers is entirely without effect in increasing the reliability of teachers' marks. If the scales themselves possess high merit, and if those who employ them do so after the best possible preparation and in the best possible manner, he believes that reliability will be increased.

APPENDIX A

THE QUESTIONS USED IN THE SCALES

CIVICS

Analysis. What are the purposes for which a political party is formed?

Cause or effect. Why is the privilege of voting very important in a republic such as ours?

Comparison. In what respects is the Constitution an improvement over the Articles of Confederation?

Criticism. Point out the strong and weak points of the initiative and referendum.

Discussion. Discuss the government of any one foreign possession of the United States.

Explanation. Explain the city manager plan.

Relationship. What is the connection between citizenship and the right to vote?

Reorganization. What are the steps in the conviction of a criminal?

Summary. Tell briefly the contents of the Constitution of the United States.

GENERAL SCIENCE

Analysis. What steps compose the process of pasteurizing milk?

Cause or effect. What causes water to rise in pumps?

Comparison. Compare surface water with deep well water for house drinking purposes.

Criticism. Criticise: "Bacteria cause many diseases among men, therefore they should be destroyed."

Discussion. Discuss the development of the Pure Food Laws.

Explanation. Tell how a siphon works.

Relationship. What is the relation of sanitation to disease?

Reorganization. Trace the history of a piece of wood from the tree to its use in a piece of furniture.

Summary. Give a short summary of the process of making steel.

AMERICAN HISTORY

Analysis. Describe the political situation connected with the Oregon boundary dispute.

Cause or effect. What was the effect of the War of 1812 upon the foreign trade of the United States?

Comparison. Compare the policies of the Republican and Democratic parties since the Civil War.

Criticism. Criticise the reconstruction policy followed after 1865.

Discussion. Discuss Civil Service reform since 1884.

Explanation. What is meant by the Monroe Doctrine?

Relationship. How was the annexation of Texas related to the slavery question?

Reorganization. What events led up to the drafting of the Monroe Doctrine?

Summary. Give a brief account of the Mexican War.

ENGLISH LITERATURE

Analysis. Show wherein the characters of *Silas Marner* grow; that is, change for the better or the worse.

Cause or effect. What effect did the coming of Eppie have upon Silas Marner?

Comparison. Compare George Eliot's method portraying character with that of Dickens.

Criticism. Criticise the character of Silas Marner.

Discussion. Discuss the place and value of the essay in literature.

Explanation. Explain the meaning of the "Romantic Movement."

Relationship. What were the relations between the Normans and the Saxons at the time of *Ivanhoe?*

Reorganization. Show in order the changes that came over Sir Launfal in his search for the Holy Grail.

Summary. State the plot of *Silas Marner* in about one hundred words.

APPENDIX B

THE RELIABILITY OF MARKING TRADITIONAL-EXAMINATION PAPERS

As was stated in a footnote on page 19, the data reported in Chapter III were computed from ratings of answers to single questions. The results from combined ratings of answers to a number of questions will be presented in this Appendix, and will be discussed especially from the standpoint of the reliability of marking pupils' responses to traditional examinations such as are commonly given by teachers. Brief summaries of two or three previous studies of the reliability of marking examination papers will also be included.

Table VIII is similar to Tables II, IV, and VI except that it contains average measures of reliability from combined ratings of pupils' answers to a number of questions, whereas the other tables contain measures from ratings of single answers. The combined ratings dealt with were secured by combining into a single paper, as it were, pupils' responses to each of the nine types of questions in each subject. In other words, it was assumed that an examination had been given in each subject consisting of nine questions, one of the analysis type, one of cause or effect, one of comparison, and so on. A pupil's score was determined by adding the ratings given by a rater to the nine answers when rated separately. Thus, since each answer was rated on a scale of ten, 90 represented a perfect score for the combined ratings.

As has been stated above, the conclusions as to the reliability of ratings with and without the scales which may be drawn from the data presented here do not differ materially from those given in Chapter III. The average coefficient of correlation for ratings with the scales is just .01 higher than for those without the scales and the average mean difference and difference in mean scores are somewhat smaller. The differences are not great enough, however, to indicate that the scales are worth using from the standpoint of increasing the reliability of rating.

Undoubtedly the best-known studies having to do with the reliability of marking ordinary examination papers are those of Starch and Elliott.[1] Of these three studies, the one dealing with mathematics has probably been referred to most often. A geometry examination consisting of ten questions of which pupils were to answer eight, with a copy of the paper written by a high-school pupil as a final exami-

[1]Starch, Daniel, and Elliott, E. C. "Reliability of the Grading of High-School Work in English," *School Review*, 20:442-57, September, 1912.
Starch, Daniel, and Elliott, E. C. "Reliability of Grading Work in Mathematics," *School Review*, 21:254-59, April, 1913.
Starch, Daniel, and Elliott, E. C. "Reliability of Grading Work in History," *School Review*, 21:676-81, December, 1913.

TABLE VIII. AVERAGE MEASURES OF RELIABILITY OF COMBINED RATINGS OF PUPILS' ANSWERS TO A NUMBER OF QUESTIONS WITHOUT AND WITH SCALES

	Coefficient of Correlation		Mean Difference		Difference in Mean Scores	
	Without	With	Without	With	Without	With
Raters						
Same	.96	.98	3.07	2.85	1.70	1.57
Different	.95	.94	7.72	7.45	5.24	5.44
Subject						
Civics	.96	.98	3.39	5.47	.94	4.60
General Science	.96	.98	5.89	5.67	2.73	4.22
History	.92	.92	6.13	4.61	5.11	2.08
Literature	.95	.96	6.63	4.63	5.43	3.35
Raters						
Same						
Participants	.96	.97	2.59	2.62	1.53	1.76
Non-Participants	.97	.98	3.85	3.16	1.99	2.24
Different						
Participants	.93	.91	9.71	6.40	9.27	3.62
Non-Participants	.95	.96	6.31	6.96	2.58	6.80
Average	.95	.96	5.56	4.95	3.59	3.36

nation, was sent to a number of North Central Association high schools with the request that the principal mathematics teacher in each school mark the paper according to the standards and practices of the school. One hundred twenty-eight usable responses were received. Those coming from schools with a passing mark of 70 varied from 25 to 89, with a probable error of 8.0; those from schools with a passing mark of 75, from 39 to 88, with a probable error of 7.2; and those from schools with a passing mark of 80, from 50 to 83, with a probable error of 7.9. Starch and Elliott also considered the question of how much of the variability found was due to the fact that the paper was marked by different teachers in different schools where the standards and practices were different. They found that the ratings of five geometry teachers in the school which the pupil attended ranged from 59 to 70, and those of four teachers in another large high school from 61 to 76. In the first case, the probable error was somewhat less than four points, and in the second case, almost five. Their conclusions from the three investigations are that marks as commonly given are extremely unreliable, and that this unreliability is "a function of the examiner and of the method of examination."

For a number of years, the studies of Starch and Elliott and others working along the same line were generally accepted as showing rather conclusively that the marks given ordinary examination papers were highly unreliable. More recently, however, some evidence has been offered by other investigators which does not support such sweeping

conclusions. Several years ago, Monroe and Souders[2] compared the reliability of teachers' marks of traditional-examination papers with that of standardized-test scores. In collecting the data for the marking of traditional examinations, two methods were used, both of which yielded data upon reliability which seem to be fairly representative of marking as actually done in regular school work. It is only fair to say, however, that they are perhaps representative of the marking done by the better teachers. The average coefficient of reliability for the results from sixty-six classes or other groups of children was found to be .65. Monroe and Souders compared this with the median of a number of coefficients of reliability for standardized tests, which was .75, and, therefore, concluded that traditional examinations may be almost as reliable as standardized tests.

Several reasons may be suggested why the difference between the reliability of ordinary marks given answers to traditional examinations and that of standardized test scores is greater than appears from a comparison of .65 and .75, but it is not necessary to enter upon a discussion of these here. The fact remains that the reliability of examination marks was found to be higher than many have assumed was the case and to compare favorably with at least some standardized tests in rather wide use.

Still more recently, Bolton[3] has presented further data on the same question. He had twenty-four arithmetic papers scored by twenty-two teachers in the same school system. An average variation of about five points[4] was found, which corresponds to a probable error of slightly more than four. About one-sixth of the variations were not greater than one, and one-third more, or one-half in all, not greater than three. Bolton concludes from these results that the marks given were on the whole decidedly satisfactory in so far as their reliability was concerned. Furthermore, he reviews one of Starch's experiments,[5] and concludes that Starch's study "seems to show great uniformity instead of great diversity as maintained by Starch." The average probable error on the ten papers employed in the study was about 4.5, but by omitting the two extreme ones it was reduced to only about 3.4. Out of the one hundred marks given (each of the ten papers was rated by ten persons) only three deviated from the average by more than 13, and all of these were given by the same instructor.

---

[2]Monroe, W. S., and Souders, L. B. "The Present Status of Written Examinations and Suggestions for Their Improvement," *University of Illinois Bulletin*, Vol. 21, No. 13, Bureau of Educational Research Bulletin No. 17. Urbana: University of Illinois, 1923, p. 27-42.

[3]Bolton, F. E. "Do Teachers' Marks Vary as Much as Supposed?" *Education*, 48:23-39, September, 1927.

[4]The ordinary percentile marking system was employed.

[5]This is not one of Starch and Elliott's studies, but one conducted by Starch alone. An account of it may be found in:
Starch, Daniel. *Educational Psychology*. New York: The Macmillan Company, 1923, p. 435.

In considering the data in Table VIII, it should be recalled that the mean differences and differences in mean scores are based on a total possible or perfect score of 90 rather than of 100 and, therefore, should be increased one-ninth to represent what they would be for the ordinary percentile marking system. If this is done, the average mean difference without the scales becomes about 6.2 and the average difference in mean scores almost exactly four. It should also be recalled that the mean differences are those between the ratings given by different raters and not those between one person's ratings and the average ratings. The probable error of these latter differences—that is, the probable error of measurement—with one-ninth added, is about 3.7. In other words, half of the ratings given by different individuals differ from the average by less than 3.7 on a scale of 100. This result agrees fairly closely with Bolton's findings.

The interpretation of variations or errors of the size just mentioned is a matter upon which not all students of examinations will agree. To some, it may seem decidedly serious, and a proof of great unreliability, that half of the differences are greater than three or four points; whereas, to others, the fact that half are smaller than this is an indication of rather satisfactory reliability. The writer takes a rather middle ground in this regard. He does not wish to be understood as claiming that traditional examinations as ordinarily given and scored, or even perhaps as given and scored under the best conditions and by the teachers most expert in their use, possess satisfactory reliability. He does, however, believe that their unreliability is not nearly so great as has been indicated by the conclusions drawn from a number of studies and as is thought by many persons. Unquestionably they are less reliable than new-type tests, whether the latter are constructed by teachers themselves or are standardized tests, but the difference is not so great as has often been suggested. Furthermore, the fact that they have less reliability is not a sufficient objection to their use to overbalance advantages which they possess for certain purposes. The general conclusion the writer would draw, therefore, is that any well-balanced testing program will include traditional examinations made and scored according to the best known technic as well as other types of measuring instruments.